Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D541–D548***
doi:10.1093/nar/gkr1130

PoSSuM: a database of similar protein–ligand binding and putative pockets

Jun-Ichi Ito[1,2], Yasuo Tabei[3], Kana Shimizu[2], Koji Tsuda[2,3] and Kentaro Tomii[1,2,*]

[1]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, [2]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064 and [3]Minato Discrete Structure Manipulation System Project, ERATO, Japan Science and Technology Agency, Sapporo 060-0814, Japan

Received August 22, 2011; Revised October 31, 2011; Accepted November 8, 2011

ABSTRACT

Numerous potential ligand-binding sites are available today, along with hundreds of thousands of known binding sites observed in the PDB. Exhaustive similarity search for such vastly numerous binding site pairs is useful to predict protein functions and to enable rapid screening of target proteins for drug design. Existing databases of ligand-binding sites offer databases of limited scale. For example, SitesBase covers only ~33 000 known binding sites. Inferring protein function and drug discovery purposes, however, demands a much more comprehensive database including known and putative-binding sites. Using a novel algorithm, we conducted a large-scale all-pairs similarity search for 1.8 million known and potential binding sites in the PDB, and discovered over 14 million similar pairs of binding sites. Here, we present the results as a relational database Pocket Similarity Search using Multiple-sketches (PoSSuM) including all the discovered pairs with annotations of various types. PoSSuM enables rapid exploration of similar binding sites among structures with different global folds as well as similar ones. Moreover, PoSSuM is useful for predicting the binding ligand for unbound structures, which provides important clues for characterizing protein structures with unclear functions. The PoSSuM database is freely available at http://possum.cbrc.jp/PoSSuM/.

INTRODUCTION

With the rapid growth in the number of solved protein structures, mainly as a result of structural genomics projects, the need for automated methods to predict protein functions from structures has become increasingly important. In many cases, proteins exhibit their biological functions by interacting with other molecules: so-called ligands. Therefore, ligand-binding sites can be regarded as functional units of the proteins. Their comparison is an important approach for elucidating protein functions. For that purpose, many methods have been developed during the last decade (1–6). To date, more than 70 000 protein structures have been available in the Protein Data Bank (PDB) (7), which contains hundreds of thousands of local regions binding to molecules of many kinds such as metal ions, nucleic acids, peptides and other small molecules. In addition to these known binding sites, numerous potential ligand-binding sites are available through various methods of ligand-binding site prediction (8–13). Performing an exhaustive similarity search for such vast quantities of protein-binding sites provides the basis for automatic classification of protein functions. Moreover, such a systematic understanding of protein–ligand interactions can be exploited for structure-based drug design.

Nevertheless, existing binding site comparison methods are applicable only to small data sets such as representative entries in the PDB, mainly because the time complexity for obtaining 3D alignment is too expensive. In many cases, the protein structure has been solved in multiple forms with a wide variety of ligands, e.g. inhibitors of different kinds. Therefore, using a representative data set results requires that the user miss such valuable samples. A reasonable strategy to retrieve similar binding sites in the PDB scale is that of employing a fast alignment-free method to detect candidates of similar binding sites first, with subsequent application of the time-consuming 3D alignment to the candidates. To achieve that objective, fingerprint-based fast similarity search approaches have been proposed recently (14–17). We have also developed an ultrafast fingerprint-based method (24) that can enumerate similar pairs from millions of binding sites in a

*To whom correspondence should be addressed. Tel: +81 3 3599 8080; Fax: +81 3 3599 8081; Email: k-tomii@aist.go.jp

reasonable time. Our method first encodes ligand-binding sites as feature vectors based on their physicochemical and geometric properties. Then similar sites are enumerated using a fast neighbor-search algorithm called SketchSort (18). For this study, we applied the proposed method to all-pair similarity searches for 1.8 million known and potential ligand-binding sites. Consequently, we discovered over 14 million pairs of similar binding sites, which is the largest-scale study of binding site comparison for the PDB entries ever reported. All the discovered similar pairs were compiled into a new database called Pocket Similarity Search using Multiple-sketches (PoSSuM). Similar binding sites have already been enumerated and stored in our database. Therefore, users can retrieve similar sites rapidly, within a few seconds, through our web interface. Because, all sites were annotated with information of various types such as CATH (19), SCOP (20), EC numbers (21) and Gene Ontology (GO) terms (22), users can easily scrutinize similar binding sites between proteins with different folds or similar catalytic sites between enzymes with different EC numbers.

In comparison with a well-known existing database, SitesBase (23), which includes ~33 000 known ligand-binding sites, our new database stores a much larger number of up-to-date known binding sites deposited in the PDB. In addition to them, our database includes pairs between known and potential ligand-binding regions predicted using a novel pocket detection program. Our database is expected to be useful for annotation of protein functions and rapid screening of target proteins in drug design. The PoSSuM database is freely available for use by researchers at http://possum.cbrc.jp/PoSSuM/.

PROCEDURE FOR DATABASE CONSTRUCTION

1.8 million binding site data set

As the source data set, we concatenated the following two sets: 241 486 known ligand-binding sites obtained from protein–ligand complexes in the PDB and 1 588 329 putative ligand-binding sites predicted using a geometric-based pocket detection algorithm.

In our study, the definition of a ligand is any HET molecule observed in the PDB (excluding water molecules, nucleic acids and peptides). The definition of a binding site is a set of amino acids around a ligand. We selected 47 562 protein entries (X-rays with resolution $\leq$4.0 Å and MODEL1 of all NMR) from the PDB (version January 2011). Of those, we found HET molecules fulfilling the definition of ligands given above. As a ligand-binding site, we extracted a set of all amino acids, each of which had at least one heavy atom lying within a distance of 5.0 Å, from at least one heavy atom of the ligand. A binding site consists of amino acids that are not necessarily continuous in the sequence and which need not be located in the same domain or chain [one example is Figure 12C in our earlier study (24)]. Consequently, we obtained 241 486 known ligand-binding sites: far more than those of SitesBase (Table 1). This difference is mainly attributable not only to the version of the PDB, but also to the ligand definition: our data set includes small molecules such as metal ions.

Table 1. Data set of protein-binding sites included in SitesBase and PoSSuM

	SitesBase	PoSSuM
PDB version	June 2005	January 2011
Definition of a ligand	HET molecule with $\geq$6 atoms	All HET molecules
Number of known ligand-binding sites (number of PDB entries)	33 168 (12 898)	241 486 (47 562)
Number of putative sites (number of PDB entries)	None	1 588 329 (29 779)

Potential binding sites were predicted using a novel binding pocket detection program: Ghecom (13). In this program, a potential ligand-binding pocket on the protein surface is detected as a region into which small probes can enter, but large probes cannot. Various shapes and sizes of pockets can be predicted by changing the radius of large probes (R_{large}). We prepared a non-redundant PDB subset (95% sequence identity cutoff and resolution $\leq$4.0 Å for X-rays) that contained 29 779 entries. For each, we iterated Ghecom four times while changing the values of R_{large} (3, 4, 5 and 6 Å) and fixing the value of R_{small} to 1.87 Å. The detected pockets are measured by the number of small probes filling up a pocket. To adjust the putative pocket size approximately to that of known ligand-binding sites, we chose pockets that were smaller than 200 probes. Then, a potential binding site was obtained by extracting a collection of all amino acids, each of which had at least one heavy atom within a distance of 5.0 Å from one of the probes. Consequently, 1 588 329 potential binding pockets were obtained (Table 1). In total, we collected 1 829 815 known and putative ligand-binding sites. Although the known sites were taken from almost all PDB entries, the putative ones account for about half of the PDB entries in the current version of PoSSuM.

Annotation to binding sites

To facilitate subsequent analyses, all binding sites were annotated to the greatest degree possible using CATH (version 3.4) codes, SCOP (version 1.75) domain classification codes EC commission numbers, and three biological domains of GO terms (molecular functions, biological processes and cellular components).

Each site was annotated with four levels of CATH codes, i.e. Class, Architecture, Topology and Homology, and of SCOP codes, i.e. Class, Fold, Superfamily and Family, by matching the binding site residues against the domain region defined by CATH or SCOP. One binding site can reside between multiple domains. In such a case, we found all domains involved with the binding site, and annotated the site with the multiple CATH and SCOP codes. Among those domains, many had not been

defined in CATH or SCOP. In our study, if the number of binding site residues that were overlapped with undefined domains was >70% of all binding site residues, then we regarded the binding site as an undefined one and assigned '0.0.0.0' to it. Furthermore, we assigned EC numbers and GO terms to binding sites. Because, an EC number and/or a GO term was assigned to each protein chain, we detected the largest overlapped protein chain for the binding site, and assigned the corresponding EC number/GO term to the site.

Fast method for enumerating similar binding sites

To enumerate similar pairs, we applied our ultrafast method to the 1.8 million binding sites. Our method encodes each binding site as a feature vector based on their physicochemical and geometric properties. Then, the similarity between two sites is measured as the cosine similarity of two vectors, ranging from −1.0 to 1.0. To compare sites, an earlier often-used method, FuzCav (15), which uses similar descriptors of binding sites to those that we use, employed a special similarity metric counting of exact matches of frequencies in the descriptors. However, this is too sensitive to small frequency changes of sites. Instead, we use cosine similarity, which is more robust against such changes, in our method (24). The cosine value drops according to the difference between two vectors, i.e. two sites. For brute-force pairwise cosine calculation, similar sites are enumerated using a fast neighbor-search algorithm called SketchSort (18), in which feature vectors are converted into bit strings and where neighbor pairs are enumerated by multiple masked sorting (25). The source code of SketchSort is available on our website (http://sites.google.com/site/tabeiyasuo/). A crucial point is that the sorting operation can be performed as approximately $O(n)$, where n represents the number of binding sites. It is much faster than a brute-force pairwise comparison whose time complexity is $O(n^2)$. The computation time to enumerate all similar pairs with cosine similarity $\geq$0.77 from the 1.8 million sites was 7 days or less on a single-core processor (Xeon 2.93 GHz; Intel Corp.). From the enumerated pairs, we eliminated the pairs that comprised only two putative sites because the pairs without a ligand label are less informative for finding ligand analogs that bind to a similar region, and for predicting the bound ligands of putative sites. We specifically examined the remaining pairs (over 277 million) consisting of at least one known ligand-binding site and performed structural alignment of all of them. To obtain the 3D superpositions for all pairs, we simply employed TM align (26). Although this program was developed to obtain 3D alignment of protein global structures, it can also be applied to two sets of discontinuous amino acids such as those of the ligand-binding site [details presented in our accepted paper (24)]. An important concern is that biologically meaningful pairs might be filtered out during this procedure because TM align can only find 3D superposition with sequence-order dependent alignment. However, we selected it mainly for the computation time because no algorithm that can perform over 277 million sequence-order independent alignments within a reasonable time has ever been reported. Eventually, we obtained 14 556 057 pairs whose binding sites were aligned with six or more residues.

Statistics of obtained pairs

Of the similar pairs, 66% (9 625 132 pairs) were known–known pairs comprising only known ligand-binding sites (Figure 1A). The remaining 34% (4 930 925) were known–putative pairs consisting of a known ligand-binding site and a putative site. These pairs, which are unique to our database, are particularly interesting because they are useful for speculating on the type of bound ligand and for assigning functions to the un-characterized proteins.

Meanwhile, 51% of detected pairs consisted of two binding sites, both of which were annotated with CATH code(s). We classified them into two categories: homologous pairs comprise two binding sites with identical CATH codes; all remaining pairs are analogous pairs. A binding site can have multiple CATH codes, as described earlier. In such a situation, if at least one pair of the CATH codes exactly matches at four levels, then the pair is regarded as homologous; otherwise it is an analogous pair. Analogous pairs were 1 973 495 (13%); homologous pairs were 5 540 245 (38%) (Figure 1B), suggesting that our method can detect similar binding sites, not only between homologous proteins but also between proteins with different folds or sequences.

All of the obtained 14 million pairs were compiled into a relational database, POSSUM, along with their corresponding annotations such as CATH, SCOP, EC and GO.

THE WEB INTERFACE

PoSSuM is implemented using PHP, MySQL and JavaScript on a machine with eight 2.93-GHz processors (core i7; Intel Corp.). One advantage of the PoSSuM database is that similar pairs have already been enumerated using our method and have been stored in the MySQL database. Consequently, similar sites can be retrieved rapidly. PoSSuM provides search modes of two types.

(i) *SearchK*: given a known ligand-binding site as a query, this search mode provides two applications. The users can retrieve various ligands whose binding sites are similar to the query site if the target data set is set to known ligand-binding sites. By searching against the putative data set, users can detect putative pockets that might bind the ligand of the query site, which might be useful for detecting apo-forms.

(ii) *SearchP*: post a known protein structure as a query, and search for similar sites for the predicted binding pockets on the query structure against known binding sites. This search mode is helpful for inferring the binding ligand of a structure of interest.

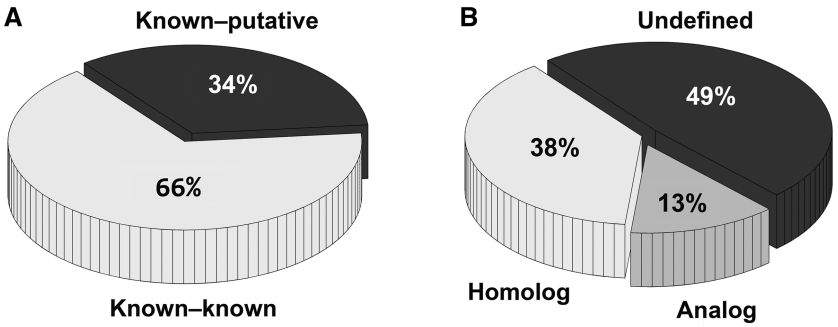


Figure 1. Categorization of 14 million obtained pairs. (**A**) 9 625 132 pairs (66%) comprised only known ligand-binding sites, whereas 4 930 925 pairs (34%) comprised a known ligand-binding site and a putative site. (**B**) Undefined pairs were 7 042 317 (49%). The remaining pairs were categorized into two groups: 5 540 245 pairs of homologs (38%) and 1 973 495 pairs of analogs (13%).

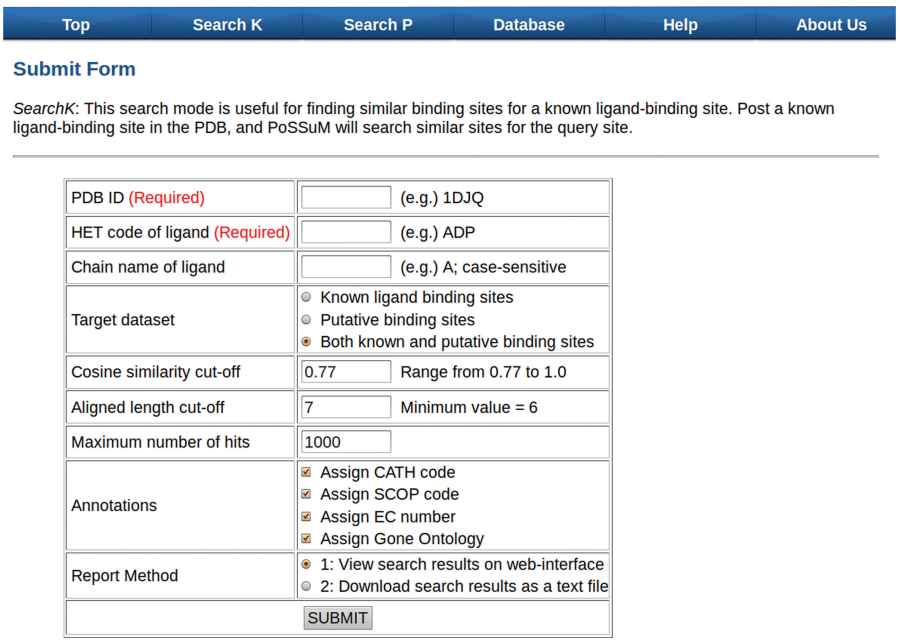


Figure 2. Submission form of PoSSuM. PDB ID and HET code of ligand should be specified for *SearchK*; PDB ID is required for *SearchP*.

Input: submit a known ligand-binding site or a whole protein structure

The required inputs for *SearchK* are PDB ID (e.g. 1DJQ) of a protein structure and the HET code (e.g. ADP, CA, or K) of a ligand bound to the structure (Figure 2). A protein structure frequently contains multiple ligands with the same HET code. Users can specify the ligand by selecting 'Chain ID'. Otherwise, all ligands with a name corresponding to the HET code the users selected are regarded as query sites. Users can select a target data set from known ligand-binding sites, putative-binding sites, or both. The only required input for *SearchP* is a PDB ID of interest. The potential binding pockets of the query structure will be searched against the known ligand-binding sites. For advanced search options such as similarity thresholds, please refer to our website.

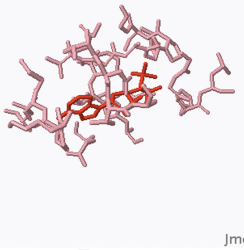
Output: report a list of similar binding sites

Once the query is posted, a summary of the query is displayed at the top of the result page (Figure 3). Sites are listed and sorted in descending order of aligned length if similar binding sites are detected against the query. Information related to each hit (similar binding site) is presented in a row in the list. The rows involved with known ligand-binding sites are shown with a blue background; putative ones are green. The first column shows the superposition of the query site to the hit. The second to fourth columns, respectively, show the PDB ID of the structure, the name of the bound ligand and the Chain ID of the ligand. Regarding putative sites, the name of the bounded ligand is simply displayed as 'PRB'. The fifth to eighth columns show similarity/dissimilarity values such as cosine similarity, *P*-value [based on our empirical

Summary of your query

Back to Submit Page

PDB ID	1DJQ
HET code	ADP
Chain name	Any
Molecule name	TRIMETHYLAMINE DEHYDROGENASE
CATH code of query site	3.40.50.720
SCOP code of query site	c.4.1.1
EC number of query chain	1.5.99.7
GO term 1: Molecular Function	GO:0016491 GO:0046872 GO:0050470 GO:0051536 GO:0051539
GO term 2: Biological Process	GO:0055114
GO term 3: Cellular Component	none
Target DB	Known and Putative binding sites
Cosine similarity cut-off	0.77
Aligned length cut-off	7
Maximum number of hits	1000
Annotates hits with CATH/SCOP/EC/GO	Yes/Yes/Yes/Yes/



Search Result

Number of hits: 1000
Number of hits: 621
Processing time to search: 0.0022590160369873(s)

Download resutls as a text file

No. of hits	PDB ID	HET code	HET_chain	Cos-value	P-value	Aligned length	RMSD(Cα)	Molecule name	CATH code	SCOP code	EC numbers	GO: Molecular Function	GO: Biological Process	GO: Cellular Component	Search again for the hit
No.1	1O94	ADP	A	0.996	0.000659757	29	0.13	TRIMETHYLAMINE DEHYDROGENASE	3.40.50.720	c.4.1.1	1.5.99.7	GO:0016491 GO:0046872 GO:0050470 GO:0051536 GO:0051539	GO:0055114	Unknown	Search it!
No.2	1O94	ADP	B	0.996	0.000661125	29	0.13	TRIMETHYLAMINE DEHYDROGENASE	3.40.50.720	c.4.1.1	1.5.99.7	GO:0016491 GO:0046872 GO:0050470 GO:0051536 GO:0051539	GO:0055114	Unknown	Search it!
No.3	1E1L	FAD	A	0.813	0.0039915	29	1.32	ADRENODOXIN REDUCTASE	3.40.50.720	c.4.1.1	1.18.1.2	GO:0004324 GO:0005515 GO:0016491 GO:0050660 GO:0050661	GO:0006629 GO:0006694 GO:0006810 GO:0008202 GO:0008203 GO:0022900 GO:0055114	GO:0005739 GO:0005759	Search it!
▪ ▪ ▪ ▪															
No.1000	2B76	FAD	A	0.914	0.00417426	27	1.5	FUMARATE REDUCTASE FLAVOPROTEIN SUBUNIT	3.50.50.60	c.3.1.4	0.0.0.0	GO:0000104 GO:0005515 GO:0009055 GO:0016491 GO:0016627 GO:0050660	GO:0006810 GO:0009061 GO:0022900 GO:0055114	GO:0016020	Search it!
No.1001	3KD9	PRB	X	0.91	0.00666289	27	2.03	COENZYME A DISULFIDE REDUCTASE	0.0.0.0	0.0.0.0	1.8.1.14	GO:0016491 GO:0050451	GO:0055114	Unknown	
No.1002	3GVC	PRB	X	0.891	0.0555578	26	3.18	PROBABLE SHORT-CHAIN TYPE DEHYDROGENASE/REDUCTASE	3.40.50.720	0.0.0.0	1.-.-.-	GO:0016491	GO:0055114	Unknown	
No.1003	1O5I	PRB	X	0.889	0.0544291	26	2.99	3-OXOACYL-(ACYL CARRIER PROTEIN) REDUCTASE	3.40.50.720	c.2.1.2	1.1.1.100	GO:0000166	GO:0055114	Unknown	

Figure 3. Result report of PoSSuM. Blue and green rows, respectively, show the known ligand-binding sites and putative sites.

study (24)], aligned length and RMSD. In addition to protein names on the ninth column, CATH, SCOP, EC numbers and GO terms that users specified in the submission form are shown. These structural and functional annotations are useful for exploring similar binding sites from different folds or from different enzymes. Only a CATH and SCOP code that accounts for the largest part of the binding site is displayed if the binding site has multiple CATH or SCOP codes. The last column provides users with a powerful function to search similar binding pockets throughout the PDB rapidly. For example, if a known ligand-binding site of interest is detected using *SearchP*, then one might then be inspired to find the binding sites that are related to the site. One could then simply click the 'Search it!' button at the last column instead of returning to the top page and resubmitting the site using *SearchK*. For further analyses, all results can also be downloaded as a plain text file, which includes a list of well-aligned amino acids whose interatomic distances of the Cα atoms are within 5.0 Å, as obtained from 3D superposition using TM align.

Superposition on JmolApplet

The first column for each hit is linked to the superposition of the hit to the query site. The superposition can be visualized using the JmolApplet (http://www.jmol.org) on a web browser. To make JmolApplet available, the Java Runtime Environment (http://www.java.com/ja/download/) should be installed on the web browser in advance. The query site and the bound ligand are shown as pink and red sticks, whereas the hit site and the bound ligand are shown as cyan and green sticks (Figure 4). For the putative sites, only the binding site is displayed with stick representation.

Demonstration and application

We next demonstrate the utility of PoSSuM database for finding ligand analogs that bind to similar protein regions

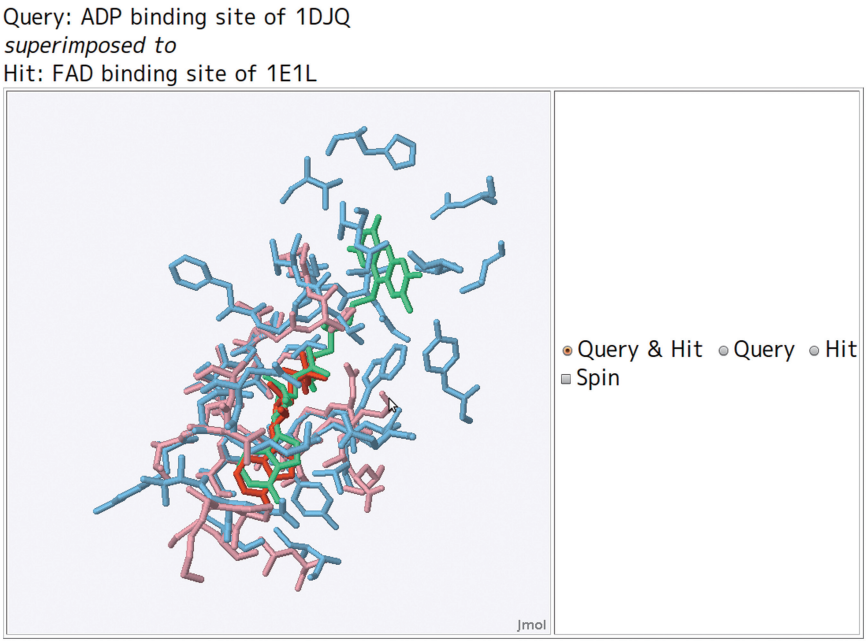


Figure 4. Color scheme for the superposed binding sites. Pink and cyan colored sticks, respectively, show the binding site of query and hit. The red sticks show a ligand of the query site (i.e. user-specified on the submit form).

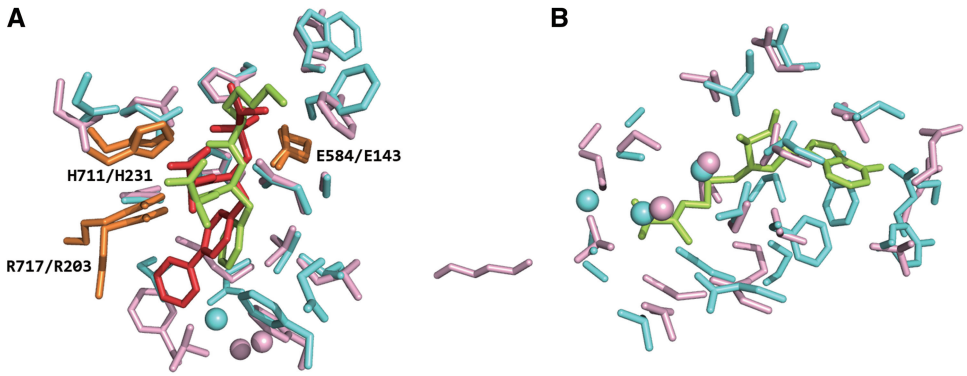


Figure 5. Superposition of a query site to a hit. (**A**) Superposition of a BIR-binding site of 1R1H (pink and red) a TI1-binding site of 1QF1 (cyan and green). Orange shows the conserved catalytic residues for both sites. (**B**) Superposition of a putative site of 3K0B (pink) and a SAH-binding site of 2ORE (cyan and green).

and the utility of assigning functions to proteins with unknown function.

In drug discovery, it is important to design chemical leads that show high specificity only to their target proteins, thereby avoiding unexpected side effects. An appropriate approach is to retrieve a set of related binding sites from non-homologous proteins. Then one can analyze the variation of ligands bound to them. *SearchK* is useful for such a task because it reports similar binding sites not only between structures within a specific family, but also among whole PDB structures. Neprilysin is involved in the regulation of amyloid β-peptide levels (27) whose structure (PDB ID: 1R1H) was determined in a complex with an inhibitor (HET code: BIR). Using the inhibitor binding site as a query, we searched for similar sites using *SearchK* and found eight similar sites: five sites are from proteins whose CATH code is identical to that of the query (i.e. 3.40.390.10), whereas three other sites are from proteins whose fold topology differs from that of the query (i.e. 3.10.170.10). Figure 5A shows superposition of an example of an analogous pair: an inhibitor (HET code: BIR) binding site from neprilysin (PDB ID: 1R1H) to an inhibitor (HET code: TI1) binding site from thermolysin (PDB ID: 1QF1). Although the global folds of the two proteins differ, their binding sites and conformations of the bound inhibitors were mutually overlapped. Furthermore, the three inhibitor-interacting residues (Glu-143, His-231 and Arg-203 of thermolysin; Glu-584, His-711 and Arg-717 of neprilysin) were well superimposed, which is regarded as representing conserved catalytic residues (28). These results support the general application of our database

for collection of various chemical compounds binding to related binding pockets. Such comprehensive information of protein–ligand interactions is expected to be useful for drug design.

To demonstrate the ability of our database to predict the type of bound ligand for unbound protein structures, an application of *SearchP* is presented below. First, we selected an unbound protein structure (PDB ID: 3K0B), as obtained from a structural genomics project, as a query for use in this demonstration. We then explored similar known ligand-binding sites for the query using *SearchP*, and retrieved 46 hits with aligned length of more than 10 residues. Of those, the top 38 hits were binding sites of *S*-adenosylmethionine (SAM) or *S*-adenosyl-L-homocysteine (SAH). The most similar hits were a SAH-binding site of 2ORE: 24 residues were aligned with RMSD 1.81 Å, as shown in Figure 5B. Furthermore, the 38 hits were from 21 different proteins whose EC numbers are limited to '2.1.1.72' and '2.1.1.-'. These results strongly indicate that the query protein structure can be bound with a *S*-adenosyl-(L)-methionine or its product *S*-adenosyl-(L)-homocysteine, and can be involved in methyltransferases (EC: 2.1.1.-), or more specifically, site-specific DNA-methyltransferase (EC: 2.1.1.72). This example demonstrates the utility of *SearchP* for predicting the type of binding ligand and function for structures whose function is unclear.

SUMMARY AND FUTURE WORK

We presented a database PoSSuM for finding similar protein–ligand binding sites. We believe that our database is a powerful tool for the annotation of protein functions and for structure-based drug design. PoSSuM will be updated as new protein structures are accumulated in the PDB. As of August 2011, PoSSuM includes over 14 million similar pairs of both known and putative-binding sites of non-polymer small molecules. We plan to include all PDB entries of proteins into our database for *SearchP* in the next update. Searching known ligand-binding sites similar to potential ones of a structure of the users' own is expected to be useful. We are developing our server so that users can upload their protein structures and analyze their potential binding sites. In addition to or instead of Ghecom, applying other method(s) to detect potential ligand-binding sites is expected to be beneficial. For example, using PocketFinder (11), unlike Ghecom, which finds potential ligand-binding envelopes rather than binding pockets on protein surface, can engender enhanced discovery of novel or hidden ligand-binding sites. We intend to improve our database in this direction. In the near future, we plan to extend our data set to binding interfaces of proteins to proteins and to nucleic acids. Performing such a comprehensive search might engender identification of overlap regions of a protein and a small molecule (29); knowledge of such regions is expected to be useful for developing inhibitors for protein–protein interaction.

ACKNOWLEDGEMENTS

We thank Masayuki Nakamura and Makoto Ishihara for technical support in developing the PoSSuM server.

FUNDING

Japan Society for the Promotion of Science (JSPS) (KAKENHI 21680025 and 23500374, partial); FIRST program. Funding for open access charge: JSPS (KAKENHI 23500374).

Conflict of interest statement. None declared.

REFERENCES

1. Schmitt,S., Kuhn,D. and Klebe,G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.
2. Shulman-Peleg,A., Nussinov,R. and Wolfson,H.J. (2004) Recognition of functional sites in protein structures. *J. Mol. Biol.*, **339**, 607–633.
3. Gold,N.D. and Jackson,R.M. (2006) Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. *J. Mol. Biol.*, **355**, 1112–1124.
4. Kinoshita,K., Murakami,Y. and Nakamura,H. (2007) eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res.*, **35**, W398–W402.
5. Kinjo,A.R. and Nakamura,H. (2007) Similarity search for local protein structures at atomic resolution by exploiting a database management system. *Biophysics*, **3**, 75–84.
6. Minai,R., Matsuo,Y., Onuki,H. and Hirota,H. (2008) Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins*, **72**, 367–381.
7. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
8. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph*, **13**, 323–330, 307–328.
9. Hendlich,M., Rippmann,F. and Barnickel,G. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph Model*, **15**, 359–363, 389.
10. Dundas,J., Ouyang,Z., Tseng,J., Binkowski,A., Turpaz,Y. and Liang,J. (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, **34**, W116–W118.
11. An,J., Totrov,M. and Abagyan,R. (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics*, **4**, 752–761.
12. Capra,J.A., Laskowski,R.A., Thornton,J.M., Singh,M. and Funkhouser,T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.
13. Kawabata,T. (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins*, **78**, 1195–1211.
14. Yin,S., Proctor,E.A., Lugovskoy,A.A. and Dokholyan,N.V. (2009) Fast screening of protein surfaces using geometric invariant fingerprints. *Proc. Natl Acad. Sci. USA*, **106**, 16622–16626.
15. Weill,N. and Rognan,D. (2010) Alignment-free ultra-high-throughput comparison of druggable protein–ligand binding sites. *J. Chem. Inf. Model*, **50**, 123–135.
16. Sciabola,S., Stanton,R.V., Mills,J.E., Flocco,M.M., Baroni,M., Cruciani,G., Perruccio,F. and Mason,J.S. (2010) High-throughput virtual screening of proteins using GRID molecular interaction fields. *J. Chem. Inf. Model*, **50**, 155–169.

17. Xiong,B., Wu,J., Burk,D.L., Xue,M., Jiang,H. and Shen,J. (2010) BSSF: a fingerprint based ultrafast binding site similarity search and function analysis server. *BMC Bioinformatics*, **11**, 47.
18. Tabei,Y., Uno,T., Sugiyama,M. and Tsuda,K. (2010) Single versus multiple sorting for all pairs similarity search, *In the 2nd Asian Conference on Machine Learning* (ACML2010), Tokyo, Japan, 13, pp. 145–160.
19. Greene,L.H., Lewis,T.E., Addou,S., Cuff,A., Dallman,T., Dibley,M., Redfern,O., Pearl,F., Nambudiry,R., Reid,A. *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.
20. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
21. Martin,A.C. (2004) PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, **20**, 986–988.
22. The GO Consortium. (2007) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
23. Gold,N.D. and Jackson,R.M. (2006) SitesBase: a database for structure-based protein–ligand binding site comparisons. *Nucleic Acids Res.*, **34**, D231–D234.
24. Ito,J., Tabei,Y., Shimizu,K., Tomii,K. and Tsuda,K. PDB-Scale analysis of known and putative ligand binding sites with structural sketches. *Proteins*, (in press).
25. Uno,T. (2010) Multi-sorting algorithm for finding pairs of similar short substrings from large-scale string data. *Knowl. Inf. Syst.*, **25**, 229–251.
26. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
27. Marr,R.A., Guan,H., Rockenstein,E., Kindy,M., Gage,F.H., Verma,I., Masliah,E. and Hersh,L.B. (2004) Neprilysin regulates amyloid Beta peptide levels. *J. Mol. Neurosci.*, **22**, 5–11.
28. Baral,P.K., Jajcanin-Jozic,N., Deller,S., Macheroux,P., Abramic,M. and Gruber,K. (2008) The first structure of dipeptidyl-peptidase III provides insight into the catalytic mechanism and mode of substrate binding. *J. Biol. Chem.*, **283**, 22316–22324.
29. Davis,F.P. and Sali,A. (2010) The overlap of small molecule and protein binding sites within families of protein structures. *PLoS Comput. Biol.*, **6**, e1000668.